Pacific Northwest
NATIONAL LABORATORY

PNNL-36742

# Molecular Vision

## Multimodal, multitask retrieval of molecular structure from measured signatures for reference-free compound identification.

September 2024

Christine Chang
Sean Colby
John Cooper
Christian Svinth
Jessie Yaros

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

## DISCLAIMER



PACIFIC NORTHWEST NATIONAL LABORATORY
*operated by*
BATTELLE
*for the*
UNITED STATES DEPARTMENT OF ENERGY
*under Contract DE-AC05-76RL01830*

**Printed in the United States of America**

**Available to DOE and DOE contractors from**
**the Office of Scientific and Technical Information,**
**P.O. Box 62, Oak Ridge, TN 37831-0062**
**www.osti.gov**
**ph: (865) 576-8401**
**fox: (865) 576-5728**
**email: reports@osti.gov**

**Available to the public from the National Technical Information Service**
**5301 Shawnee Rd., Alexandria, VA 22312**
**ph: (800) 553-NTIS (6847)**
**or (703) 605-6000**
**email: info@ntis.gov**
**Online ordering: http://www.ntis.gov**

PNNL-36742

# Molecular Vision

Multimodal, multitask retrieval of molecular structure from measured signatures for reference-free compound identification

September 2024

Christine Chang
Sean Colby
John Cooper
Christian Svinth
Jessie Yaros

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

# Abstract

We are currently at risk of generating false conclusions based on limited methods to identify small molecules in biological systems and in chemical forensics. By definition, the chemical structures of novel small molecules have not been determined, let alone measured or synthesized. Currently, unambiguous structure determination of small molecules is constrained by the time and effort needed to isolate compounds and perform de novo structure elucidation using laboratory-based methods, significantly extending the time to inform mitigation strategies. To address this gap, we have developed a deep learning approach to directly map molecular structure to experimental signatures. We aim to unify measurement technologies employed in untargeted small molecule identification studies—such as infrared (IR) spectrometry, tandem mass spectrometry (MS/MS), ion mobility spectrometry-derived collision cross section (CCS)—through use of a multimodal, multitask deep learning architecture. Where existing methods require direct generation of information-rich spectra and/or properties, an inherently difficult task, we will simplify molecular signature-based identification by posing the problem as a recognition or retrieval task. The model is thus presented with relevant endpoints – structure and one or more molecular signatures – and need only determine whether they are semantically related. Thus, our approach offers the following advantages over existing techniques: (i) circumvents difficulties associated with direct generation of molecular signatures from structure and structure from signatures; (ii) incorporates multiple molecular signatures simultaneously, as available, to support identification; and (iii) enables rapid computation of structural embeddings toward broad coverage of known chemical space. Taken together, the approach removes the need to explicitly obtain or compute reference spectra, representing a powerful method for compound identification that requires only experimentally observed signatures.

# Summary

We developed an approach to perform molecular structure retrieval from multiple measurement sources: IR and MS/MS. Framing molecular identification in this manner eliminates the need for reference libraries containing measured experimental nor computationally predicted signatures. Our approach utilized state-of-the-art networks to embed molecular structure and corresponding signatures, relating proximity in the embedding space according to the Tanimoto similarity between endpoints. An advanced loss function, InfoNCE, was ultimately utilized to maximize mutual information among like pairs, theoretically resulting in rich embedded representations. Together, these selections culminated in high validation accuracy of 90.48% and 78.91% for structure:IR and structure:MS/MS, respectively. However, such pairwise assessments do not generalize to real-world performance, as evaluated by top-$k$ accuracy: even our most performant model was only able to achieve 0.397% top-5 accuracy.

# Acknowledgments

This research was supported by the *m/q* Initiative, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). A portion of the research was performed using resources available through Research Computing at PNNL. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

# Acronyms and Abbreviations

CCS: collision cross section
NMR: nuclear magnetic resonance
IR: infrared
MS/MS: tandem mass spectrometry
GNN: graph neural network
nD: one, two,..n-dimensional
CNN: convolutional neural network
InfoNCE: information noise contrastive estimation
NIST: National Institute of Standards and Technology
EPA: Environmental Protection Agency
JCAMP-DX: Joint Committee for Atomic and Molecular Physical Data Exchange
GNPS: Global Natural Product Social Molecular Networking
SMILES: simplified molecular line input system
InChI: international chemical identifier
CAS: Chemical Abstract Service
*m/z*: mass-to-charge ratio
CID: compound identification
CLERMS: contrastive learning-based embedder for the representation of tandem mass spectra
SELU: scaled exponential linear unit
GLDM: graph latent diffusion model
FiLMConv: feature-wise linear modulation
RELU: rectified linear unit
MoLeR: molecule level reward
BCE: binary cross entropy
L2: Euclidean distance
PyPI: Python Package Index
AMD: Advanced Micro Devices
DDR4: double data rate fourth generation
RAM: random access memory
CUDA: compute unified device architecture
MHz: megahertz
GB: gigabyte
HBM2: high bandwidth memory second generation
GPU: graphics processing unit

# Contents



# Figures

# Equations

# 1.0 Introduction

A major challenge in the identification of molecules via high-throughput metabolomics studies is a lack of reference data against which to query. Experimental signatures are typically searched against reference libraries containing spectra and properties from analyses of authentic reference standards. However, chemical reference libraries currently represent less than 1% of known molecules, require significant time and resources to expand, and do not include compounds that are difficult to obtain or synthesize.[1] While researchers have demonstrated nascent success with *in silico* prediction of some molecular attributes, such as collision cross section (CCS),[2,3] nuclear magnetic resonance (NMR) chemical shifts,[4] infrared (IR) spectra,[5] and tandem mass spectra fragmentation patterns (MS/MS),[6,7] limitations in throughput and accuracy have stymied broad, immediate use. Moreover, existing approaches have yet to fully leverage multiple measurements simultaneously to inform probabilistic annotations.

We hypothesize that the lackluster performance of *in silico* prediction methods are testament to the challenge of modeling the relevant dynamics under complex conditions. Existing methods require direct generation of information-rich spectra and/or properties, an inherently difficult task. Instead, molecular signature-based identification can be posed as a recognition or retrieval task,[8] as opposed to the more complex prediction task. The model is thus presented with relevant endpoints – structure and one or more molecular signatures – and need only determine whether they are *semantically* related.

To this end, we have developed a deep learning architecture to learn a joint representation of molecular structure and experimental signatures. By employing methods widely used in the image-text retrieval space[9]—that is, given an image, retrieve relevant descriptive text (and vice versa)—we jointly encoded a shared representation of molecular graphs and their associated molecular signatures. To build this joint embedding, we implemented domain specific embedding networks: a graph neural network (GNN)[10,1112] to embed a 3D molecular structure and (one-dimensional) 1D convolutional neural networks (CNNs)[13] to embed corresponding signatures. We trained the network by optimizing over an information noise contrastive estimation (InfoNCE) loss[14] modified by the inclusion of a Tanimoto similarity objective. We trained the network to simultaneously minimize the error with respect to Tanimoto similarity between embedding pairs and to maximize the similarity of positive embedding pairs arising from the same structure. In other words, the distance between joint embeddings was conditioned such that matching structure-signature pairs will coalesce in the embedded vector space, whereas incorrect pairs will diverge, and allows for ambiguity according to structural similarity.

# 2.0 Methods

## 2.1 Data and Preprocessing

### 2.1.1 Infrared (IR) Spectral Data

Experimental IR spectra were obtained from the National Institute of Standards and Technology (NIST) Standard Reference Database 35[15] (henceforth, NIST35), containing both NIST and Environmental Protection Agency (EPA) values. The dataset was comprised of 5,228 spectra in the Joint Committee for Atomic and Molecular Physical Data Exchange (JCAMP-DX) format. Spectra from NIST were reported in the 550 - 3846 $cm^{-1}$ range, while EPA values ranged from 450 - 3966 $cm^{-1}$, both at 4 $cm^{-1}$ resolution. To achieve uniformity, values were truncated to the overlapping 550 – 3846 $cm^{-1}$ range, linearly interpolated to share wave number, and normalized such that vectors were of unit norm.

### 2.1.2 Tandem Mass Spectrometry (MS/MS) Data

Tandem mass spectra were obtained from the NIST20 tandem mass spectral library, Global Natural Product Social Molecular Networking (GNPS) aggregated database, and RIKEN.[16–19] Tandem mass spectral data corresponding to NIST35 compounds were observed in both GNPS and NIST20, amounting to 15,441 GNPS spectra and 20,961 NIST20 spectra, representing 1,085 (21%) of compounds with associated IR data. We additionally included the remainder of NIST20 spectra for which compounds had a valid Chemical Abstract Service (CAS) lookup number, totaling 952,336 spectra representing 24,102 compounds. To remove the potential confounding effect of including the precursor mass-to-charge ratio (*m/z*) in tandem mass spectra, according to,[20] we additionally removed all fragments with *m/z* greater than or equal to the precursor *m/z*. Tandem mass spectra, both with and without precursor *m/z*, were evaluated.

### 2.1.3 Collision Cross-Section (CCS) Data

Experimentally acquired CCS values were compiled from the McLean CCS Compendium[21] and CCSbase.[22] Note that the values from CCSbase are sourced from multiple data sources between 2014–2022.[22–45] In total, 14,854 CCS values for 8,690 compounds are represented in the full database. The overlap of CCS values with either IR or MS/MS spectra was surprisingly small, at an intersection of 1000 compounds. For this reason, and due to difficulties achieving success with the more information-rich IR and MS/MS spectra, CCS was not evaluated for the purposes of this report.

### 2.1.4 Molecular Structures

Molecular structures were sourced as two-dimensional (2D) graphs from NIST35 in MDL Molfile format (.mol), indexed by CAS number. Missing structures were reconciled by CAS identifier lookup using the cactus Chemical Identifier Resolver (cactus.nci.nih.gov). For MS/MS and CCS databases (NIST20, GNPS, RIKEN, McLean, and CCSbase), available identifiers, including SMILES (simplified molecular line input system),[46] InChI (international chemical identifier),[47] CAS registry number, and PubChem Compound Identification (CID), were used to determine structure. Each identifier, as available, was used to lookup or otherwise compute, using the RDKit chemistry toolbox (www.rdkit.org), a representation by SMILES. All SMILES were canonicalized using RDKit to ensure uniformity. To ensure a consistent index of molecular identity, canonicalized SMILES were converted to InChI hash keys, or InChI key, and used to index compounds henceforth.

Additionally, to enable structural similarity comparison during training, the molecular fingerprint[48] of each molecule was calculated using RDKit using default *RDKitFingerprint*. Parameters were specified as $minPath = 1$, $maxPath = 7$, $fpSize = 2048$, $bitsPerHash = 2$, $useHs = True$, $tgtDensity = 0.0$, and $minSize = 128$. During training, Tanimoto similarity[48] is evaluated for the structures relevant to each modal endpoint according to their precomputed fingerprints.

## 2.2 Infrared Embedding Model

To embed infrared signals, we explore using 1D deep residual networks,[49] 1D Xception networks,[50] and 1D Vision Transformers.[51,52]

### 2.2.1 ResNet

Deep residual networks,[49] or ResNets, are a type of convolutional network typically used in computer vision tasks to provide a performance baseline to more advanced models. They benefit from having small computational overhead due to their limited complexity, allowing them to be scaled to arbitrary depths. Moreover, their convolutional architecture provides a consistent interpretability across layers in the form of low, middle, and high level feature maps.

To apply ResNets to signal tasks, it is simple enough to convert all two-dimensional convolution filters to one-dimensional convolution filters. Several instances of the success of ResNets have been recorded in human bioactivity and industrial monitoring tasks,[53–55] as well as in mass spectra classification tasks,[56] collision cross-section prediction,[57] and mass spectra annotation.[58] Specific to our task is the application of one-dimensional convolutional neural network variants to infrared embedding and classification tasks.[59–61] Specifically, Zhang et al.[62] introduce the idea of blending the features from the Inception Network[63] with a basic ResNet, which suppresses the resolution of deeper convolutional filters while maintaining performance and reducing computational overhead. With the noted successes of ResNet variants across domains, and applied to various types of temporal and non-temporal data, we choose to apply ResNet as an embedding network for infrared spectra.

### 2.2.2 Xception

A major difference between the Xception network[50] - specifically its depth-wise separable convolutions - and a regular convolutional neural network is that the prior proposes for cross-channel correlations and spatial correlations to be learned entirely independent of one another. In Xception's depth-wise separable convolutions, a set of channel-wise spatial convolutions, typically $3\times3$, are applied over an image or signal, producing a separate convolution for each channel. Then a point-wise $1 \times 1$ convolution is applied across all channels output from the earlier spatial convolutions. By following this sequence of convolutions, it is hypothesized that spatial information (channel-wise spatial convolutions) and cross-channel information (point-wise convolution) can be learned separately, thereby unburdening the network from having to account for these simultaneously.[50] Such modifications to the convolutional architecture have led to increases in accuracy and model efficiency.[50] To use the Xception network on infrared signals, we similarly replace all two-dimensional convolutions with one-dimensional convolutions and apply no other changes. To our knowledge, there are no instances of a direct application of the Xception network to infrared signals. While the 1D Xception variation was explored, we found the ResNet implementation more performant, and thus used that variation in our experimentation.

### 2.2.3 Vision Transformer

Vision transformers were introduced by Dosovitskiy et al. as a method to exploit the base transformer[64] architecture for computer vision tasks.[52] The main difference between these

architectures is the use of convolution layers in the prior to encode image patches which are then fed into the virtually unmodified architecture of the latter. In similar fashion to converting a ResNet to work for one-dimensional applications, one can simply replace the two-dimensional convolutions in Vision Transformers with one-dimensional convolutions for use in spectral applications. There are, however, few instances of 1D vision transformers being applied to signal data in the literature. Weng et al., for example, apply one-dimensional vision transformers with multi-scale convolutions for bearing fault diagnosis,[65] while Dong et al. apply one-dimensional vision transformers with deformable convolutions for arrhythmia classification.[65] To our knowledge there are no instances of 1D vision transformers applied directly to infrared or near-infrared spectra. We amend a PyTorch implementation of a one-dimensional vision transformer, but replace the dense layers in the patch embedding with one-dimensional convolutions for better performance.

## 2.3 Tandem Mass Spectra Embedding Model

We experiment with several embedding models for MS/MS. Unlike infrared spectra, whose intensities are collected at a constant interval, MS/MS spectra include a variable number of $m/z$–intensity pairs corresponding to the fragmentation pattern of a specific precursor. As a result, the models used to encode[1] them necessarily reflect these differences. Specifically, special care is taken in all MS/MS-embedding models to project variable-length MS/MS sequences bijectively into an $nD$ space.

MS/MS spectra are typically encoded by choosing a fixed number of bins, and therefore a fixed resolution window, to group $m/z$ values and their associated peaks. By discretizing the real $m/z$ space, subsequent analysis forgoes the distinguishing, high-resolution information captured in the $m/z$ peaks neighbored within a small interval. Such relevant peak information is inevitably collapsed by and conflated within the binning process.[66] For this reason, in all MS/MS-embedding models, we choose to replace discretized binning with a bijective mapping between peaks and fixed-length numerical peak vectors following the methods outlined in Voronov et al.[66] and explicated below in Equation 1.

$$\mathrm{SE}(m/z, 2i; d) = \sin\left(2\pi\left[\lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2i/(d-2)}\right]^{-1} m/z\right)$$

$$\mathrm{SE}(m/z, 2i+1; d) = \cos\left(2\pi\left[\lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2i/(d-2)}\right]^{-1} m/z\right)$$

Equation 1. Sinusoidal embedding.

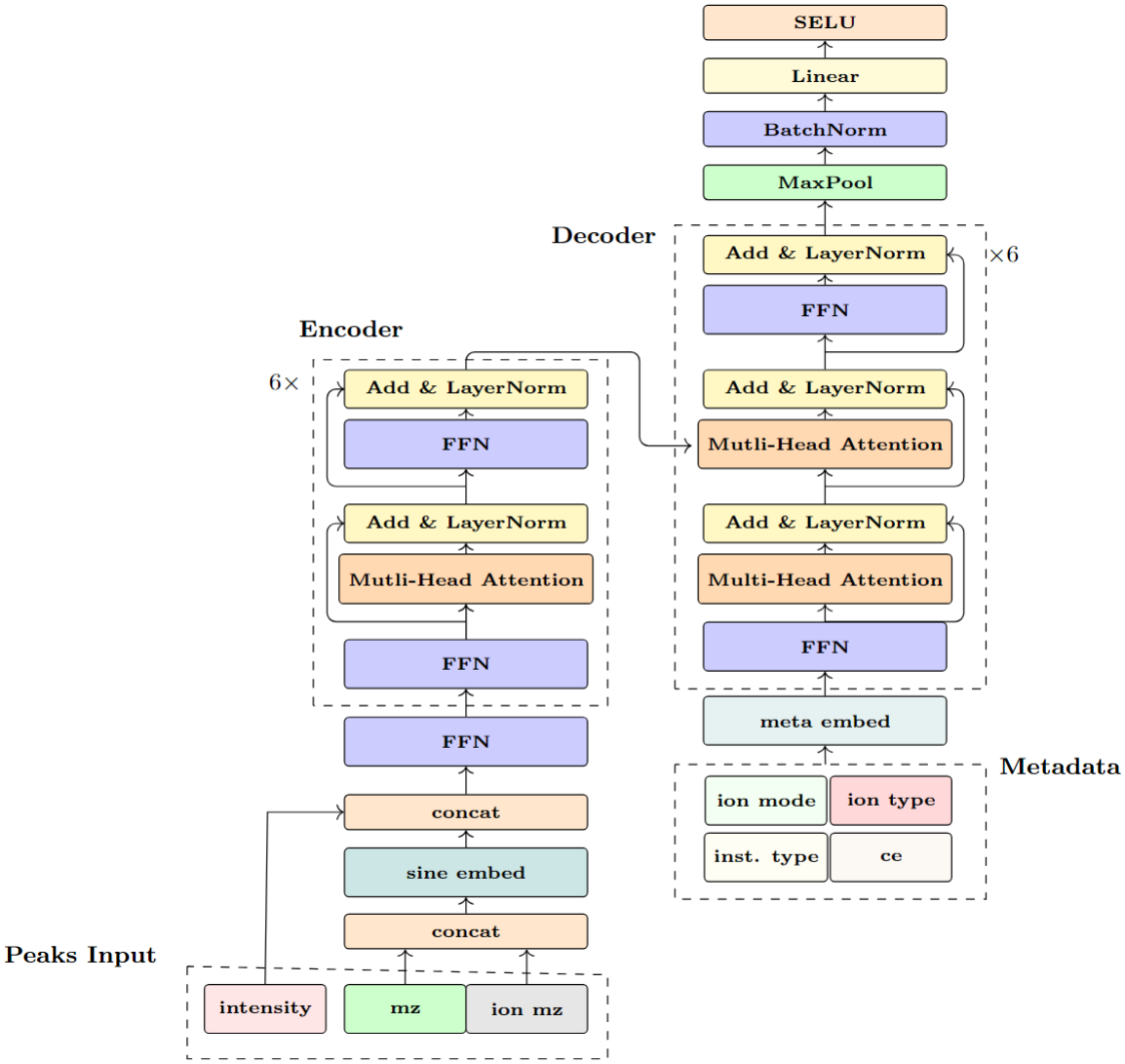


Figure 1.Transformer architecture for MS/MS embeddings.

To embed MS/MS signals, we explore using plain transformer networks outfitted with the sinusoidal encodings described in Voronov et al. as contrastive learning-based embedder for the representation of tandem mass spectra (CLERMS).[67] The model architecture consists of six encoding and six decoding layers. For input, the original precursor mass is corrected according to adduct type, producing a shifted neutral molecule mass which we label as *ion m/z*, and whose derivation is seen in Equation 2.

$$f(x)_{\text{Molecule m/z}\rightarrow\text{Ion m/z}} = \begin{cases} x + 1.0072764665789 & adduct(x) = [\text{M+H+}] \\ x + 22.989222 & adduct(x) = [\text{M+Na}]+ \\ x - 1.0072764665789 & adduct(x) = [\text{M-H}]- \\ x - 34.969402 & adduct(x) = [\text{M+Cl}]- \end{cases}$$

Equation 2. Mapping between neutral molecule *m/z* and adduct ion *m/z*.

To ensure uniformity of input length, and roughly following the methods outlined by Guo et al.,[67] MS/MS spectra are selected for their hundred highest peaks or padded with zeros if one hundred peaks aren't available. The MS/MS spectra and ion *m/z* are then concatenated and fed into a sinusoidal embedding module which encodes the spectra via Equation 1. Intensity is then concatenated to this encoding, and it is passed through a series of dense layers before entering the transformer encoder's layers. This output is then fed to the cross-attention head in the decoder.

Along with the encoder output, we encode the ionization mode, ionization type, instrument type, and collision energy meta data of the spectra by passing them through separately initialized embedding networks each comprising a small feed forward network. These embeddings are then concatenated and fed into the decoder along with the encoder output. The decoder output is then projected to a 512-dimension vector, and sent through a $MaxPooling$, $BatchNorm$, and $Linear$ layer before a scaled exponential linear unit ($SELU$) activation sends it to its final representation.

## 2.4 Structure Embedding Model

### 2.4.1 AttentiveFP

To facilitate the acquisition of robust structural molecular encodings by the multimodal model, molecules were represented as graph structures suitable for neural network consumption. This process, referred to as molecular featurization, requires the creation of node vectors characterizing individual atoms and edge vectors representing bonds between pairs of atoms. Key attributes embedded within node vectors encompass atom type, formal charge, hybridization, and the number of hydrogen bonds, while edge vectors incorporate information including bond order, whether the atoms within the pair are part of a ring, and the conjugation status of the bond.

These features were then passed into AttentiveFP, a GNN designed for molecular property prediction tasks.[68] AttentiveFP leverages attention mechanisms to discern non-local relationships between atoms, encouraging the model to extract the most salient aspects of molecular structure essential for accurately mapping molecules to their corresponding experimental signatures.

Together the featurization and GNN components constitute the structural encoder within our multimodal architecture. To implement this process, molecular geometries in MDL Molfile format were loaded using RDKit. Next, these molecular structures underwent featurization utilizing the

MolGraphConvFeaturizer from the DeepChem library.[69] The featurized molecular graphs were then fed into the AttentiveFP model from the DGL-LifeSci package.[70] Notably, the AttentiveFP implementation required a minor modification to yield intermediate embeddings of desired size, rather than scalar predictions.

### 2.4.2 GLDM

As an additional method for encoding molecular structures, we leveraged the pre-trained graph encoding network from the encoder portion of the graph latent diffusion model (GLDM).[71] The GLDM encoder network is a molecular graph variational autoencoder which has been pre-trained on the GuacaMol[72] dataset for the task of encoding molecular graphs into a latent space via the encoding portion of the network, and then reconstructing the original graph from the latent representation via the decoding network. For our work, we solely leverage the encoding portion of the network. We fine-tuned the pre-network on our training dataset of molecular graphs following the training hyper-parameters outlined in the implementation of the GLDM. The overall architecture of the encoder is a graph neural network, composed of 12 GNN blocks consisting of a FiLM convolutional[73] layer, a normalization layer, and a rectified linear unit (ReLU)[74] layer.

During pre-training, the GLDM autoencoder is trained according to the methodology described by molecule level reward (MoLeR),[75] a paper which approaches the problem of molecule generation by decomposing larger molecules into 'motifs' which are then combined to create full molecules. We use the MoLeR algorithm to extract motif features from our dataset, then use the extracted features to represent the molecules in our dataset as PyTorch Geometric[76] graphs for ingestion by the pre-trained model. In this graph formulation, nodes are motifs (either individual atoms or small structures of atoms) and edges are bonds between the atoms/structures.

## 2.5 Multimodal Retrieval Network

Each mode-specific embedding network necessarily produces an equal-length vector representation of its input, enabling direct comparability among modes. Thus, for a particular modal pair, corresponding embedding networks are engaged in the forward pass to produce respective embeddings, as seen in Figure 3. The resulting vectors are then compared, depending on the particular loss selected, according to a measure that confers the similarity between them. Each loss type explored during the project follows.

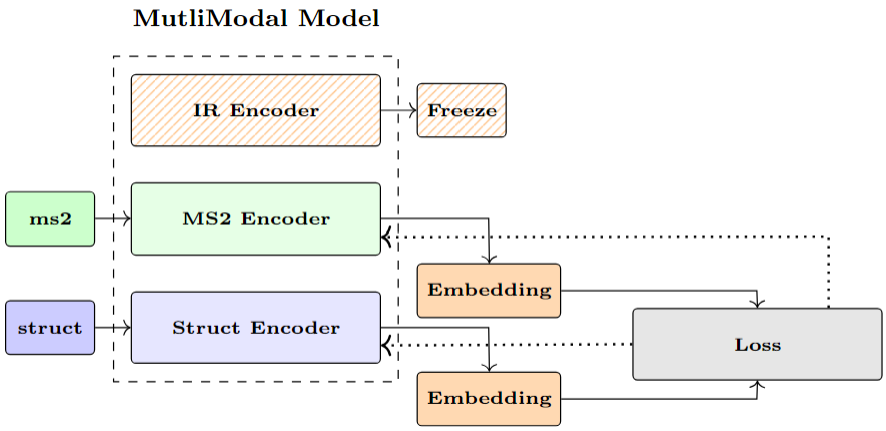


Figure 2. Multimodal network single-pair input flow. Dotted lines indicate back-propagation.

### 2.5.1 Binary Cross Entropy Loss (BCE)

The simplest form of loss was achieved by alternating positive (same) and negative (different) training instances, per modal pair. In this framing, we explored use of several distance metrics: cosine, Euclidean distance (L2), and a "learned" distance metric, whereby the network was configured with an additional layer to predict the distance metric among instances. Labels were initially supplied as strictly positive (label = 1) or negative (label = 0). We later modified this configuration to support the natural ambiguity among similar, but not identical, structures, by way of Tanimoto similarity. Thus, for example, cases in which two similar structures would otherwise be given a binary label of strictly 0, Tanimoto similarity allowed for intermediate values between 0 and 1.

### 2.5.2 Contrastive Loss

While BCE loss focuses on the correctness of the final binary decision ("same" versus "different"), contrastive loss directly optimizes the relative distances between modal pairs, which is critical for understanding and comparing complex patterns in signals. Therefore, contrastive loss more naturally aligns with the goal of distinguishing between similar and dissimilar signal pairs, often leading to superior performance in such tasks.

We implemented contrastive loss with optional terms for positive margin – the distance at which like pairs are not further *concentrated* in the embedding space – and negative margin – the distance at which different pairs are not further *dispersed* in the embedding space.

### 2.5.3 InfoNCE Loss

A variant of contrastive loss, but sufficiently different to warrant its own section, is information noise contrastive estimation (InfoNCE) loss.[14] With InfoNCE, all possible pairings within a particular batch, i.e. all $N$ positive pairs and $N(N-1)$ negative pairs, are considered for the gradient update. This framing maximizes use of the already-embedded modal representations per batch, conferring maximum information to the update step. Additionally, we modified the loss term employed by CLERMS[67] for MS/MS retrieval: instead of separate terms for InfoNCE and mean squared error against Tanimoto similarity, we pulled the Tanimoto target inside the InfoNCE evaluation, creating a unified loss term. These considerations maximize mutual information between positive pairs, encouraging the model to learn richer representations, and improves scalability, relative to the pairwise comparisons of standard contrastive loss.

This information is captured in Equation 3. Here, $Q$ and $R$ denote paired inputs with shared molecular structure. Direct comparisons of these pairs comprise the positive examples in our batch. All other pairwise comparisons comprise negative examples, i.e. inputs not inheriting from a shared structure. Our loss maximizes positive similarity between positive pairs, while simultaneously encouraging negative pairs to converge to their pre-computed Tanimoto similarity, denoted as *S*.

$$\mathrm{L}(Q,R,S) = \sum_{i=0}^{K} -\log\left(\frac{\exp(q_i \cdot r_i)}{\sum_{j=0}^{K}\exp(\mathrm{abs}(q_i \cdot r_j - S_{ij}))}\right), \text{with}\left\{(q_i,r_j)\;\middle|\;\begin{array}{l}(q_i,r_j)\in Q\times R\\ S(q_i,r_i)=1\\ S(q_i,r_j)\neq 1\end{array}\right\}$$

Equation 3. Modified InfoNCE loss that considers pairwise Tanimoto similarity.

## 2.6 Data Loader

Equally important as the neural network architecture, a well-architected PyTorch Dataset object acts as an interface between data and model, allowing efficient and convenient access to training instances. Critically, it enables the loading and preprocessing of data before consumption by the neural network such as normalization, resizing, interpolation, etc., ensuring that the data is in a suitable format for training. Additionally, the Dataset object helps optimize memory usage by dynamically loading and processing data in batches, minimizing the memory footprint and enabling training on large datasets that may not fit entirely in memory. In the context of training a multimodal retrieval network, where a multitude of possible input pairings are associated with indication as to their correspondence, the Dataset object plays a crucial role in brokering the data-model relationship.

Along these lines, class balance can be controlled, which is critical in multimodal retrieval network contexts. For a training set of size $N$ with $M$ modes, there are $N(M-1)$ possible positively labeled pairings and $N(N-1)(M-1)/2$ possible negatively labeled pairings. If strictly randomly sampled, the positive class would be severely underrepresented, leading to predictions biased to the negative class.

During training, each batch alternates among possible mode pairings (e.g. structure:IR, structure:MS/MS, IR:MS/MS, etc.). For BCE and contrastive loss, positive and negative instance pairs are sampled randomly with equal proportion. Thus, for each epoch, all possible positive instance pairs are seen, alongside an equal number of randomly sampled negative instance pairs. For InfoNCE loss, positive and negative pairs are constructed from a batch of $N$ InChI keys: pairwise comparisons are made among all batch members, where the diagonal represents "same" pairs, and the off-diagonal elements "different" pairs.

## 2.7 Training

Spectral data sourced from NIST20[16,17] was split into train (70%), validation (15%), and test (15%) partitions by molecular identifier (InChI key) to ensure no leakage between sets. The network was trained to minimize selected loss term and was optimized using the Adam optimizer[77] with $5 \times 10^{-6}$ learning rate and an exponential learning rate scheduler with $\gamma = 0.999$. Training was performed with a batch size of 64 for 300 epochs. When using BCE loss and/or contrastive loss in a binary setting, validation metrics including $AUPRC$, $AUROC$, and accuracy were calculated following each epoch across all modal pairs, but also for all pairwise modalities used. In other words, both general and modality specific metrics were tracked during the training process. When using InfoNCE loss, only validation loss was tracked. The best-performing model iteration with respect to validation loss was kept.

To train a multi-modal network with different modality-specific sub-networks, only those sub-networks whose respective modalities were being used during the training step were allowed to receive gradient updates, while the remaining subnetwork's weights were frozen (Figure 3). This remained the default setting for training runs unless another paradigm was directly specified (such as a fully frozen structural encoder).

## 2.8 Package Implementation

The multimodal retrieval network implementation, training and validation scripts, and example Jupyter notebooks have been made available as a Python package: *molvis*. We architected *molvis* to adhere to software development best practices, including installation through Anaconda or the Python Package Index (PyPI), in-line documentation via docstrings, and version control with Git. Upon achieving success and subsequent publication, the *molvis*

package will be open-source and freely available online at github.com/pnnl/molvis, and community contributions via pull request will be welcome.

## 2.9 Compute Resources

Computations were performed on the Pacific Northwest National Laboratory Research Computing cluster, Deception, comprised of 188 compute nodes, each with 64 cores (dual AMD EPYC 7502 processors at 2.5 GHz) and 256 GB DDR4 RAM. For machine learning training and evaluation, nodes equipped with Nvidia Tesla V100 (12 nm lithography, 5120 CUDA cores at 1246 MHz, 32 GB HBM2 memory) GPUs were utilized.

# 3.0 Results and Discussion

Initial results with AttentiveFP[68] (structure), ResNet[49] (IR), and CLERMS[67] (MS/MS) subnetworks, trained with BCE loss, appeared optimistic: structure:IR validation accuracy of 72.7% and structure:MS/MS validation accuracy of 69.0%. However, the context of this result is important, as this represents the proportion of *pairwise* comparisons that were judged correctly by the network, *not* the ability of the network to discern the correct structure from a (ostensibly large) library.

Real world retrieval performance must be evaluated using another metric, separate from the task the network was trained on: typically, top-$k$ accuracy. That is, in what proportion of queries does the correct classification appear in the $k$ top-scoring outputs. For molecule identification, this would ideally mean high top-1 accuracy, where the top-scoring model output contains the correct classification. Our well-performing network with respect to binary accuracy produced surprisingly low top-$k$ results. Even with $k = 5$, accuracy remained below 1%.

This led to the exploration of loss functions better equipped to handle the intricacies of inter-embedding distances across modal pairs. First among them, contrastive margin loss, which quickly netted improvements to binary validation accuracy: 76.6% and 72.6% for structure:MS/MS and structure:IR, respectively. Other small improvements, such as batchbalancing, learning rate modifications, and contrastive hinge adjustments eventually yielded a maximum binary validation accuracy of 90.48% and 78.91% for structure:IR and structure:MS/MS. Despite these improvements, both in theoretical soundness of the approach and according to pairwise accuracy assessments, top-$k$ remained sub-1%. Use of InfoNCE, hypothesized to further improve embedding space conditioning, also yielded a sub-1% top-$k$ accuracy.

Following this, we conducted unimodal spectral retrieval experiments for MS/MS:MS/MS and IR:IR. Achieving passable results, we speculated that capacity of the structural encoder was insufficient for the multimodal task. We therefore decided to freeze the rich pretrained GLDM[71] structural encoder and lower the learning rate to $1 \times 10^{-7}$ in order to see if we could move our spectral embeddings in the direction of GLDM's structural embeddings. This again yielded a sub-1% top-$k$ accuracy.

We hypothesize that the poor top-$k$ performance could be due to one or more of the following factors: (i) Insufficient weight in the loss term for positive pairs. The current configuration weights batch size $N$ positive pairs equal to $N^2$ negative pairs, potentially resulting in overfitting latent space dispersion of negatives at the expense of positive coalescence. (ii) Modality competition, generally. Huang et al. suggest that modal representations which correlate more with the randomly initialized weights of an encoding network will be better learned than those which do not.[78] Although their experiment uses a single encoding network for multimodal training, the idea of modality-biased model weight shifting is still an open concern in our experiments. (iii) Lacking a unified, fused representation for modal pairs. Contrastive, BCE, and InfoNCE[14] losses seek to push similar embeddings together and disparate embeddings apart. However, these optimizing strategies are not forced to create a cohesive, single-vector representation for inputs.

# 4.0 Conclusion

We developed an approach to perform molecular structure retrieval from multiple measurement sources: IR and MS/MS. Framing molecular identification in this manner eliminates the need for reference libraries containing measured experimental nor computationally predicted signatures. Our approach utilized state-of-the-art networks to embed molecular structure and corresponding signatures, relating proximity in the embedding space according to the Tanimoto similarity between endpoints. An advanced loss function, InfoNCE, was ultimately utilized to maximize mutual information among like pairs, theoretically resulting in rich embedded representations. Together, these selections culminated in high validation accuracy of 90.48% and 78.91% for structure:IR and structure:MS/MS, respectively. However, such pairwise assessments do not generalize to real-world performance, as evaluated by top-$k$ accuracy: even our most performant model was only able to achieve 0.397% top-5 accuracy.

# 5.0 References


(1) Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; others Non-target screening with highresolution mass spectrometry: critical review using a collaborative trial on water analysis. *Analytical and bioanalytical chemistry* 2015, *407*, 6237–6255.

(2) Colby, S. M.; Thomas, D. G.; Nuñez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; others ISiCLE: a quantum chemistry pipeline for establishing in silico collision cross section libraries. *Analytical chemistry* 2019, *91*, 4346–4356.

(3) Colby, S. M.; Nuñez, J. R.; Hodas, N. O.; Corley, C. D.; Renslow, R. R. Deep learning to generate in silico chemical property libraries and candidate molecules for small molecule identification in complex samples. *Analytical chemistry* 2019, *92*, 1720–1729.

(4) Yesiltepe, Y.; Nuñez, J. R.; Colby, S. M.; Thomas, D. G.; Borkum, M. I.; Reardon, P. N.; Washton, N. M.; Metz, T. O.; Teeguarden, J. G.; Govind, N.; others An automated framework for NMR chemical shift calculations of small organic molecules. *Journal of cheminformatics* 2018, *10*, 1–16.

(5) Henschel, H.; Andersson, A. T.; Jespers, W.; Mehdi Ghahremanpour, M.; van der Spoel, D. Theoretical Infrared Spectra: Quantitative Similarity Measures and Force Fields. *Journal of Chemical Theory and Computation* 2020, *16*, 3307–3315, Publisher: American Chemical Society.

(6) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Analytical chemistry* 2021, *93*, 11692–11700.

(7) Koopman, J.; Grimme, S. Calculation of mass spectra with the QCxMS method for negatively and multiply charged molecules. *Journal of the American Society for Mass Spectrometry* 2022, *33*, 2226–2242.

(8) Gan, Z.; Li, L.; Li, C.; Wang, L.; Liu, Z.; Gao, J. Vision-Language Pre-training: Basics, Recent Advances, and Future Trends. 2022, Number: arXiv:2210.09263 arXiv:2210.09263 [cs].

(9) Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; Huang, J. Vision-Language Pre-Training with Triple Contrastive Learning. 2022, Number: arXiv:2202.10401 arXiv:2202.10401 [cs].

(10) Choudhary, K.; DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials* 2021, *7*, 1–8, Number: 1 Publisher: Nature Publishing Group.

(11) Zhang, S.; Liu, Y.; Xie, L. Molecular Mechanics-Driven Graph Neural Network with Multiplex Graph for Molecular Structures. 2020, Number: arXiv:2011.07457 arXiv:2011.07457 [physics, q-bio].

(12) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* 2018, *148*, 241722, Publisher: American Institute of Physics.

(13) Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D. J. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing* 2021, *151*, 107398.

(14) Oord, A. v. d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* 2018,

(15) Stein, S. E. NIST 35. NIST/EPA Gas-Phase Infrared Database - JCAMP Format. *NIST* 2008, Last Modified: 2008-10-16T10:10-04:00 Publisher: Stephen E. Stein.

(16) Yang, X.; Neta, P.; Stein, S. E. Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra. *Analytical Chemistry* 2014, *86*, 6393–6400.
(17) Yang, X.; Neta, P.; Stein, S. E. Extending a Tandem Mass Spectral Library to Include MS 2 Spectra of Fragment Ions Produced In-Source and MS n Spectra. *Journal of the American Society for Mass Spectrometry* 2017, *28*, 2280–2287.
(18) Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* 2016, *34*, 828–837.
(19) Tsugawa, H. et al. A lipidome atlas in MS-DIAL 4. *Nature Biotechnology* 2020,
(20) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature Methods* 2021, *18*, 1524–1531.
(21) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Collision cross section compendium to annotate and predict multi-omic compound identities. *Chemical Science* 2019, *10*, 983–993.
(22) Ross, D. H.; Seguin, R. P.; Krinsky, A. M.; Xu, L. High-Throughput Measurement and Machine Learning-Based Prediction of Collision Cross Sections for Drugs and Drug Metabolites. *Journal of the American Society for Mass Spectrometry* 2022, *33*, 1061–1072.
(23) May, J. C.; Goodwin, C. R.; Lareau, N. M.; Leaptrot, K. L.; Morris, C. B.; Kurulugama, R. T.; Mordehai, A.; Klein, C.; Barry, W.; Darland, E.; Overney, G.; Imatani, K.; Stafford, G. C.; Fjeldsted, J. C.; McLean, J. A. Conformational Ordering of Biomolecules in the Gas Phase: Nitrogen Collision Cross Sections Measured on a Prototype High Resolution Drift Tube Ion Mobility-Mass Spectrometer. *Analytical Chemistry* 2014, *86*, 2107–2116, PMID: 24446877.
(24) Paglia, G.; Williams, J. P.; Menikarachchi, L.; Thompson, J. W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J.; Palsson, B. O.; Astarita, G. Ion Mobility Derived Collision Cross Sections to Support Metabolomics Applications. *Analytical Chemistry* 2014, *86*, 3985–3993, PMID: 24640936.
(25) Groessl, M.; Graf, S.; Knochenmuss, R. High resolution ion mobility-mass spectrometry for separation and identification of isomeric lipids. *Analyst* 2015, *140*, 6904–6911.
(26) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z.-J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Analytical Chemistry* 2016, *88*, 11084–11091, PMID: 27768289.
(27) Bijlsma, L.; Bade, R.; Celma, A.; Mullin, L.; Cleland, G.; Stead, S.; Hernandez, F.; Sancho, J. V. Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Analytical Chemistry* 2017, *89*, 6583–6589, PMID: 28541664.
(28) Hines, K. M.; Herron, J.; Xu, L. Assessment of altered lipid homeostasis by HILICion mobility-mass spectrometry-based lipidomics. *Journal of Lipid Research* 2017, *58*, 809–819.
(29) Hines, K. M.; Ross, D. H.; Davidson, K. L.; Bush, M. F.; Xu, L. Large-Scale Structural Characterization of Drug and Drug-Like Compounds by High-Throughput Ion MobilityMass Spectrometry. *Analytical Chemistry* 2017, *89*, 9023–9030, PMID: 28764324.
(30) Hines, K. M.; Waalkes, A.; Penewit, K.; Holmes, E. A.; Salipante, S. J.; Werth, B. J.; Xu, L. Characterization of the Mechanisms of Daptomycin Resistance among GramPositive Bacterial Pathogens by Multidimensional Lipidomics. *mSphere* 2017, *2*, 10.1128/msphere.00492–17.
(31) Stow, S. M.; Causon, T. J.; Zheng, X.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; Hann, S.; Fjeldsted, J. C. An

Interlaboratory Evaluation of Drift Tube Ion Mobility–Mass Spectrometry Collision Cross Section Measurements. *Analytical Chemistry* 2017, *89*, 9048–9055.

(32) Zheng, X.; Aly, N. A.; Zhou, Y.; Dupuis, K. T.; Bilbao, A.; Paurus, V. L.; Orton, D. J.; Wilson, R.; Payne, S. H.; Smith, R. D.; Baker, E. S. A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry. *Chem. Sci.* 2017, *8*, 7724–7736.

(33) Zhou, Z.; Tu, J.; Xiong, X.; Shen, X.; Zhu, Z.-J. LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility–Mass Spectrometry-Based Lipidomics. *Analytical Chemistry* 2017, *89*, 9559–9566.

(34) Blaženović, I.; Shen, T.; Mehta, S. S.; Kind, T.; Ji, J.; Piparo, M.; Cacciola, F.; Mondello, L.; Fiehn, O. Increasing Compound Identification Rates in Untargeted Lipidomics Research with Liquid Chromatography Drift Time–Ion Mobility Mass Spectrometry. *Analytical Chemistry* 2018, *90*, 10758–10764.

(35) Lian, R.; Zhang, F.; Zhang, Y.; Wu, Z.; Ye, H.; Ni, C.; Lv, X.; Guo, Y. Ion mobility derived collision cross section as an additional measure to support the rapid analysis of abused drugs and toxic compounds using electrospray ion mobility time-of-flight mass spectrometry. *Anal. Methods* 2018, *10*, 749–756.

(36) Mollerup, C. B.; Mardal, M.; Dalsgaard, P. W.; Linnet, K.; Barron, L. P. Prediction of collision cross section and retention time for broad scope screening in gradient reversedphase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. *Journal of Chromatography A* 2018, *1542*, 82–88.

(37) Nichols, C. M.; Dodds, J. N.; Rose, B. S.; Picache, J. A.; Morris, C. B.; Codreanu, S. G.; May, J. C.; Sherrod, S. D.; McLean, J. A. Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility-Mass Spectrometry. *Analytical Chemistry* 2018, *90*, 14484–14492.

(38) Righetti, L.; Bergmann, A.; Galaverna, G.; Rolfsson, O.; Paglia, G.; Dall'Asta, C. Ion mobility-derived collision cross section database: Application to mycotoxin analysis. *Analytica Chimica Acta* 2018, *1014*, 50–57.

(39) Tejada-Casado, C.; Hernández-Mesa, M.; Monteau, F.; Lara, F. J.; del Olmo-Iruela, M.; García-Campaña, A. M.; Le Bizec, B.; Dervilly-Pinel, G. Collision cross section (CCS) as a complementary parameter to characterize human and veterinary drugs. *Analytica Chimica Acta* 2018, *1043*, 52–63.

(40) Hines, K. M.; Xu, L. Lipidomic consequences of phospholipid synthesis defects in Escherichia coli revealed by HILIC-ion mobility-mass spectrometry. *Chemistry and Physics of Lipids* 2019, *219*, 15–22.

(41) Leaptrot, K. L.; May, J. C.; Dodds, J. N.; McLean, J. A. Ion mobility conformational lipid atlas for high confidence lipidomics. *Nature Communications* 2019, *10*, 985.

(42) Dodds, J. N.; Hopkins, Z. R.; Knappe, D. R. U.; Baker, E. S. Rapid Characterization of Per- and Polyfluoroalkyl Substances (PFAS) by Ion Mobility Spectrometry–Mass Spectrometry (IMS-MS). *Analytical Chemistry* 2020, *92*, 4427–4435.

(43) Poland, J. C.; Leaptrot, K. L.; Sherrod, S. D.; Flynn, C. R.; McLean, J. A. Collision Cross Section Conformational Analyses of Bile Acids via Ion Mobility–Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* 2020, *31*, 1625–1631.

(44) Tsugawa, H. et al. MS-DIAL 4: accelerating lipidomics using an MS/MS, CCS, and retention time atlas. *bioRxiv* 2020,

(45) Vasilopoulou, C. G.; Sulek, K.; Brunner, A.-D.; Meitei, N. S.; Schweiger-Hufnagel, U.; Meyer, S. W.; Barsch, A.; Mann, M.; Meier, F. Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. *Nature Communications* 2020, *11*, 331.

(46) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 1988, *28*, 31–36.
(47) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of cheminformatics* 2015, *7*, 1–34.
(48) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry* 2004, *2*, 3204–3218.
(49) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2015, arXiv:1512.03385 [cs].
(50) Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017, arXiv:1610.02357 [cs].
(51) Zhang, T.; Chen, S.; Wulamu, A.; Guo, X.; Li, Q.; Zheng, H. TransG-Net: Transformer and Graph Neural Network Based Multi-Modal Data Fusion Network for Molecular Properties Prediction. *Applied Intelligence* 2022, *53*, 16077–16088.
(52) Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* 2020, *abs/2010.11929*.
(53) Dong, L.; Wang, C.; Yang, G.; Huang, Z.; Zhang, Z.; Li, C. An Improved ResNet-1d with Channel Attention for Tool Wear Monitor in Smart Manufacturing. *Sensors* 2023, *23*.
(54) Sánchez-Reolid, R.; López de la Rosa, F.; López, M. T.; Fernández-Caballero, A. Onedimensional convolutional neural networks for low/high arousal classification from electrodermal activity. *Biomedical Signal Processing and Control* 2022, *71*, 103203.
(55) Yoo, J.; Yoo, I.; Youn, I.; Kim, S.-M.; Yu, R.; Kim, K.; Kim, K.; Lee, S.-B. Residual onedimensional convolutional neural network for neuromuscular disorder classification from needle electromyography signals with explainability. *Computer Methods and Programs in Biomedicine* 2022, *226*, 107079.
(56) Seddiki, K.; Saudemont, P.; Precioso, F.; Ogrinc, N.; Wisztorski, M.; Salzet, M.; Fournier, I.; Droit, A. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nature Communications* 2020, *11*.
(57) Samukhina, Y.; Matyushin, D.; Grinevich, O.; Buryak, A. A Deep Convolutional Neural Network for Prediction of Peptide Collision Cross Sections in Ion Mobility Spectrometry. *Biomolecules* 2021, *2021*, 1904.
(58) Kudriavtseva, P.; Kashkinov, M.; Kertész-Farkas, A. Deep Convolutional Neural Networks Help Scoring Tandem Mass Spectrometry Data in Database-Searching Approaches. *Journal of proteome research* 2021,
(59) Tan, A.; Wang, Y.; Zhao, Y.; Zuo, Y. 1D-inception-resnet for NIR quantitative analysis and its transferability between different spectrometers. *Infrared Physics & Technology* 2023, *129*, 104559.
(60) Jiang, D.; Qi, G.; Hu, G.; Mazur, N.; Zhu, Z.; Wang, D. A residual neural network based method for the classification of tobacco cultivation regions using near-infrared spectroscopy sensors. *Infrared Physics & Technology* 2020, *111*, 103494.
(61) Zhang, X.; Li, Y.; Tao, Y.; Wang, Y.; Xu, C.; Lu, Y. A novel method based on infrared spectroscopic inception-resnet networks for the detection of the major fish allergen parvalbumin. *Food Chemistry* 2021, *337*, 127986.
(62) Zhang, Y.; Davison, B. D. Modified distribution alignment for domain adaptation with pre-trained Inception ResNet. *arXiv preprint arXiv:1904.02322* 2019,

(63) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *CoRR* 2014, *abs/1409.4842*.
(64) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *CoRR* 2017, *abs/1706.03762*.
(65) Weng, C.; Lu, B.; Yao, J. A One-Dimensional Vision Transformer with Multiscale Convolution Fusion for Bearing Fault Diagnosis. 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing). 2021; pp 1–6.
(66) Voronov, G.; Lightheart, R.; Davison, J.; Krettler, C. A.; Healey, D.; Butler, T. Multiscale Sinusoidal Embeddings Enable Learning on High Resolution Mass Spectrometry Data. 2023.
(67) Guo, H.; Xue, K.; Sun, H.; Jiang, W.; Pu, S. Contrastive Learning-Based Embedder for the Representation of Tandem Mass Spectra. *Analytical Chemistry* 2023, *95*, 7888– 7896, PMID: 37172113.
(68) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; others Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* 2019, *63*, 8749– 8760.
(69) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; https://www.amazon.com/ Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.
(70) Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; Karypis, G. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS omega* 2021, *6*, 27233–27238.
(71) Wang, C.; Ong, H. H.; Chiba, S.; Rajapakse, J. C. GLDM: hit molecule generation with constrained graph latent diffusion model. *Briefings in Bioinformatics* 2024, *25*, bbae142.
(72) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* 2019, *59*, 1096–1108.
(73) Brockschmidt, M. Gnn-film: Graph neural networks with feature-wise linear modulation. International Conference on Machine Learning. 2020; pp 1144–1152.
(74) Nair, V.; Hinton, G. E. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10). 2010; pp 807–814.
(75) Maziarz, K.; Jackson-Flux, H.; Cameron, P.; Sirockin, F.; Schneider, N.; Stiefl, N.; Segler, M.; Brockschmidt, M. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864* 2021,
(76) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds. 2019.
(77) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 2014,
(78) Huang, Y.; Lin, J.; Zhou, C.; Yang, H.; Huang, L. Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably). 2022; https://arxiv.org/abs/2203.12221.

# Appendix A

This page intentionally left blank.

# Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

***www.pnnl.gov***